SPLASH: Simple yet Effective Node Property Prediction on Edge Streams under Distribution Shifts

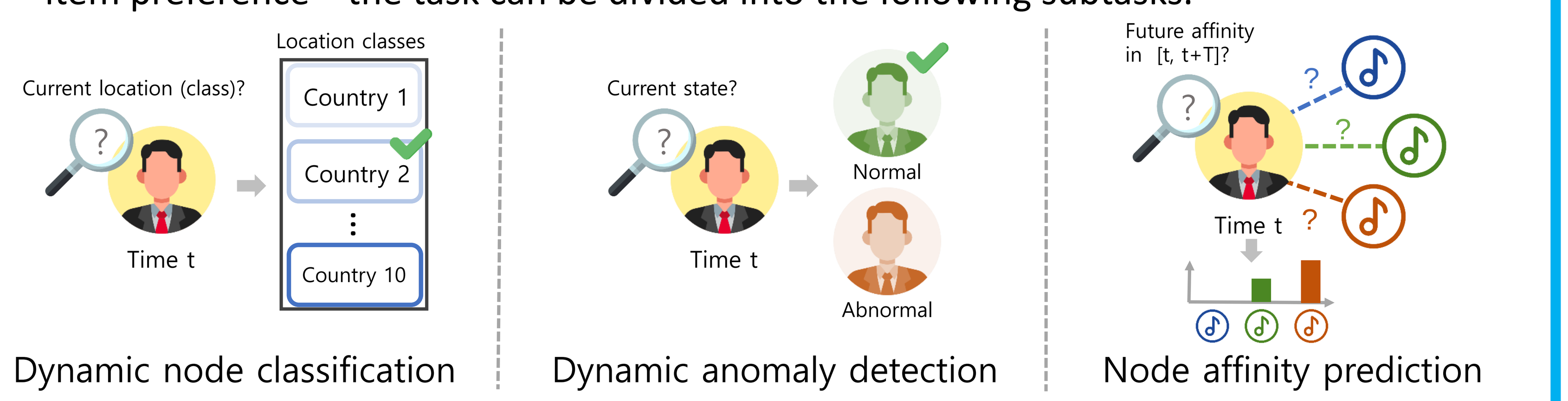
Jongha Lee — KAIST AI — jhsk777@kaist.ac.kr
Taehyung Kwon — KAIST AI — Taehyung.kwon@kaist.ac.kr
Heechan Moon — KAIST AI — heechan9801@kaist.ac.kr
Kijung Shin — KAIST AI — kijungs@kaist.ac.kr

Summary

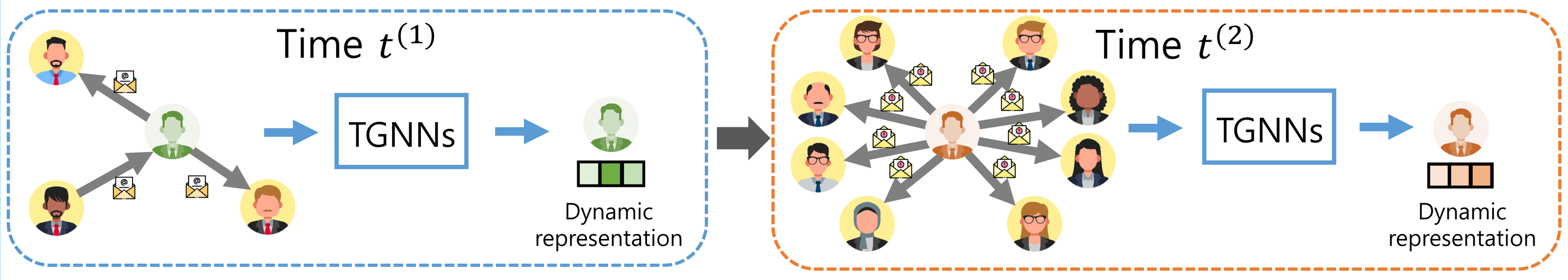
- **Motivation**
  - To address two limitations (performance degradation when node features are absent & complex model architecture) of Temporal Graph Neural Networks (TGNNs) in node property prediction under distribution shifts, which are common in real-world scenarios.
- **Proposed Method: SPLASH**
  - We propose SPLASH for efficient and effective prediction of node properties in edge streams under distribution shifts.
  - SPLASH introduces (1) node feature augmentation processes on edge streams, (2) a task-specific feature selection scheme, and (3) SLIM, a simple TGNN model that leverages them.
- **Contribution**
  - **Fast & Lightweight**: SPLASH uses only MLP layers, enabling fast inference.
  - **Effective**: SPLASH outperforms baselines in node property prediction.
  - **Robust**: SPLASH shows the smallest performance drop as the distribution shift intensifies.

Introduction

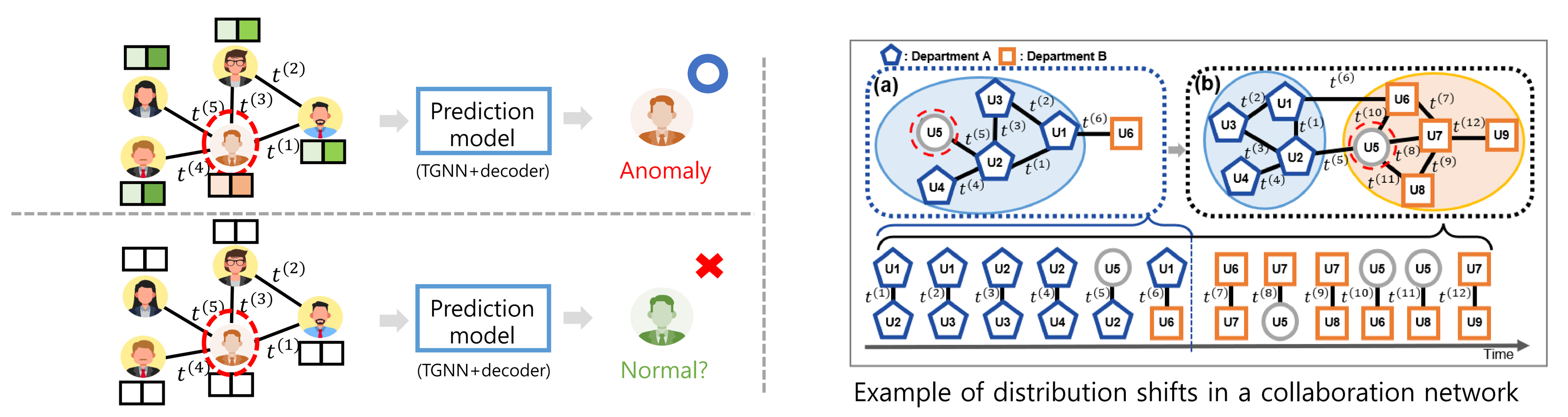
- **Node Property Prediction in Real-world Applications**
  - **What are node properties?**: Various characteristics of entities are often represented as node properties, and predicting missing properties has many real-world applications.
  - **Challenge:** node properties can change dynamically in real-world networks.
  - **Subtasks**: Depending on the target property—user class, anomalous state, or short-term item preference—the task can be divided into the following subtasks.



- **Temporal Graph Neural Networks (TGNNs)**
  - **Definition:** TGNNs are neural networks designed for evolving graphs (edge streams) by incrementally updating node embeddings that capture temporal and structural patterns over time, enabling predictions on dynamically changing node properties.



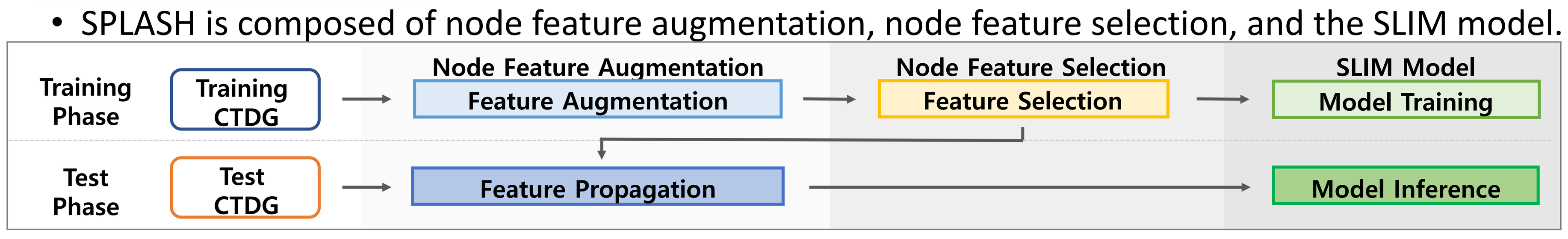
- **Limitations of TGNNs in Node Property Prediction**
  - **Limitation 1:** TGNNs are less effective when node features are absent. However, in real-world graphs, it is often difficult to obtain additional external node features.
  - **Limitation 2:** TGNNs are vulnerable to distribution shifts, common in dynamic real-world graphs, due to their complex architectures such as RNNs and attention modules.



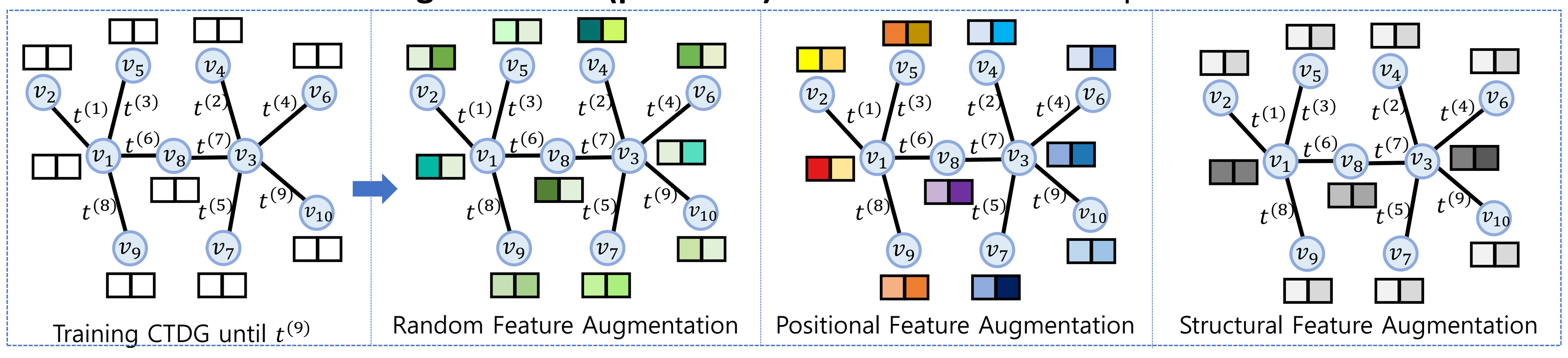
Example of distribution shifts in a collaboration network

Proposed Method: SPLASH

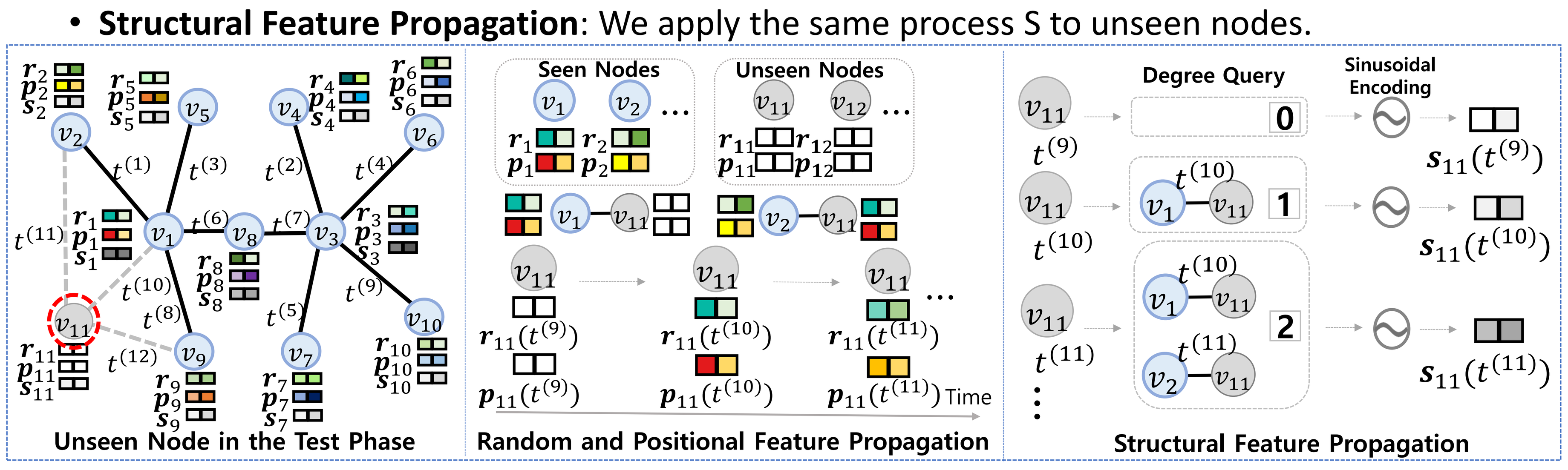
- **Overview of SPLASH**
  - SPLASH is composed of node feature augmentation, node feature selection, and the SLIM model.



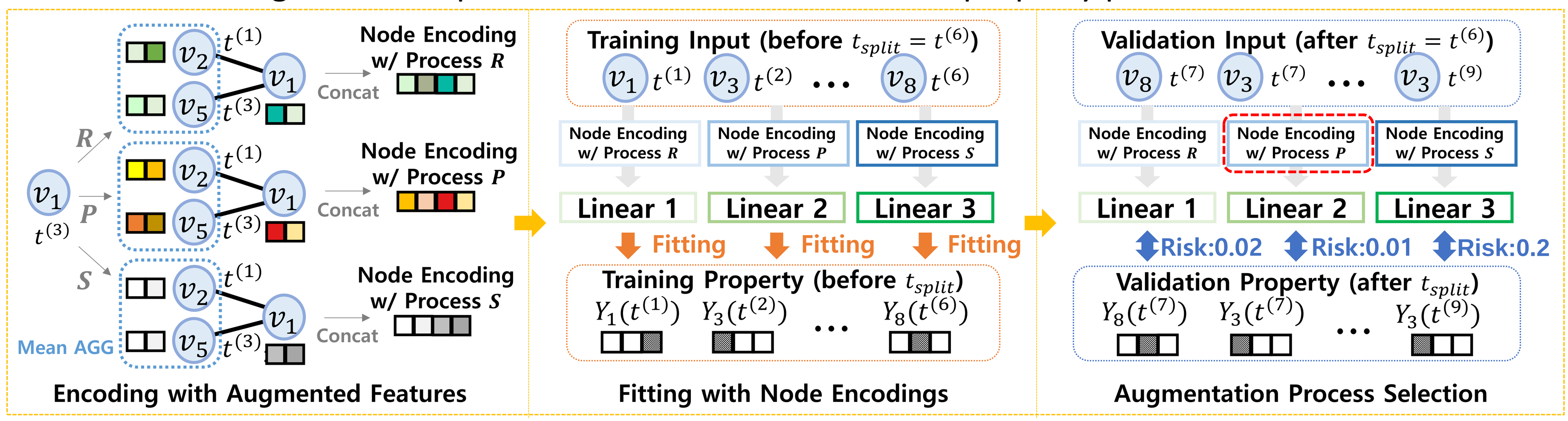
- **Node Feature Augmentation – Feature Augmentation (Training Phase)**
  - **Random Feature Augmentation (process R)** - To encode absolute positions of seen nodes.
  - **Positional Feature Augmentation (process P)** - To encode relative positions of seen nodes.
  - **Structural Feature Augmentation (process S)** - To encode structural patterns of seen nodes.



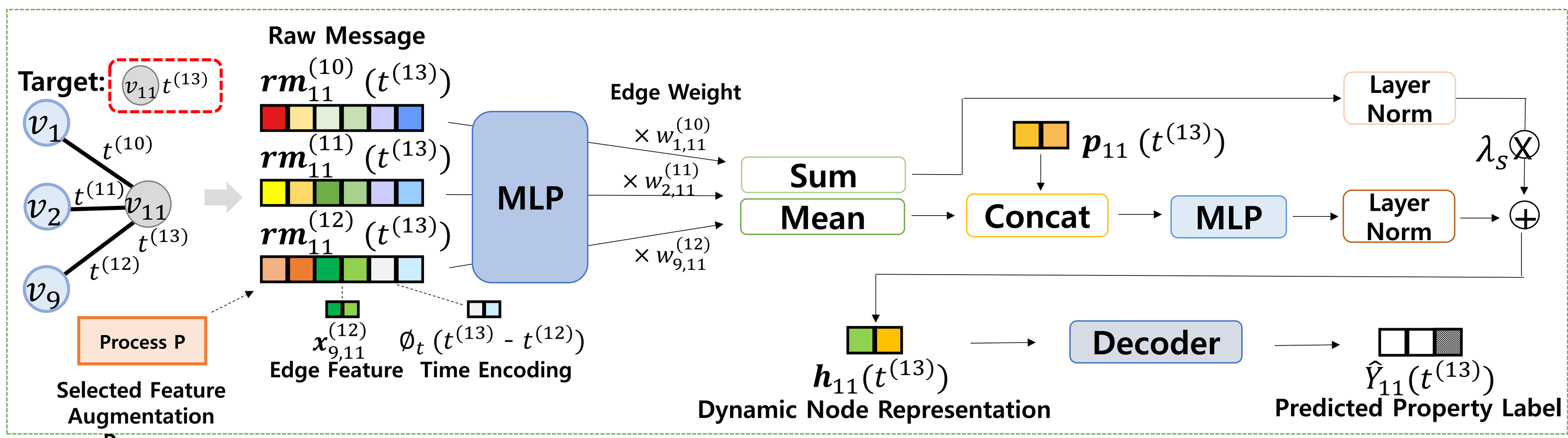
- **Node Feature Augmentation – Feature Propagation (Test Phase)**
  - **Random and Positional Feature Propagation**: We obtain the node features of unseen nodes using a simple linear interpolation of neighboring nodes' features.
  - **Structural Feature Propagation**: We apply the same process S to unseen nodes.



- **Node Feature Selection (Training Phase)**
  - **Feature Selection** is performed in the training phase using linear layers to efficiently identify a feature augmentation process that is effective for node property prediction.

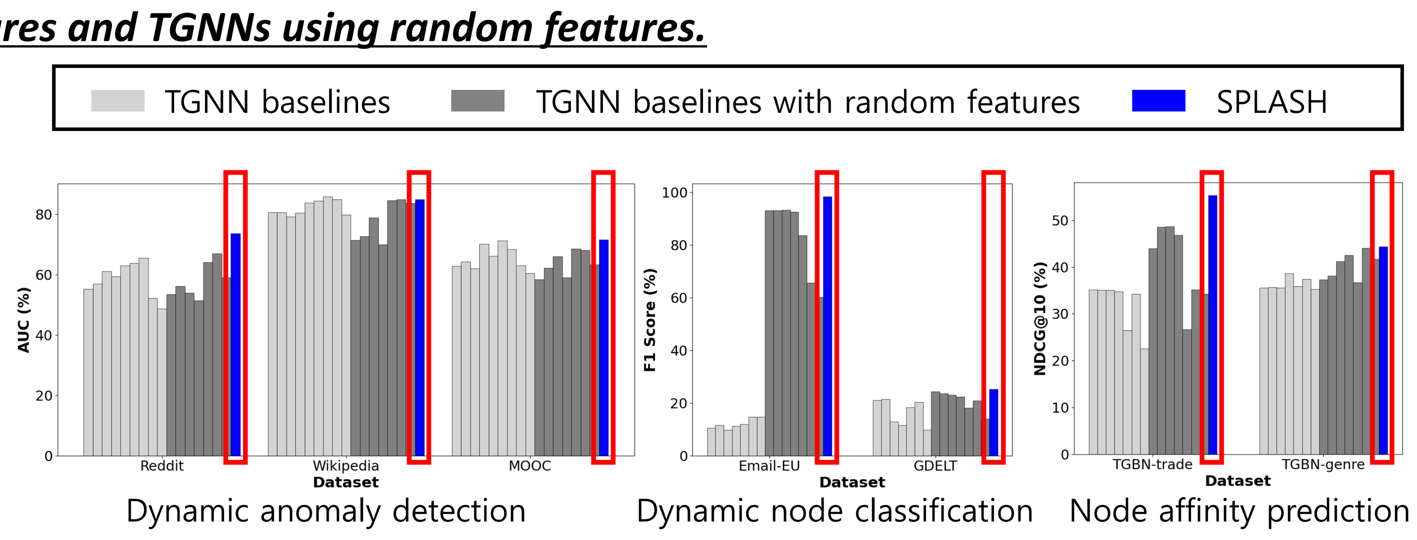


- **SLIM Model (Both Training and Test Phases)**
  - **The SLIM model** is our proposed TGNN model to efficiently and effectively predict node properties using a simple MLP-based model with augmented node features.

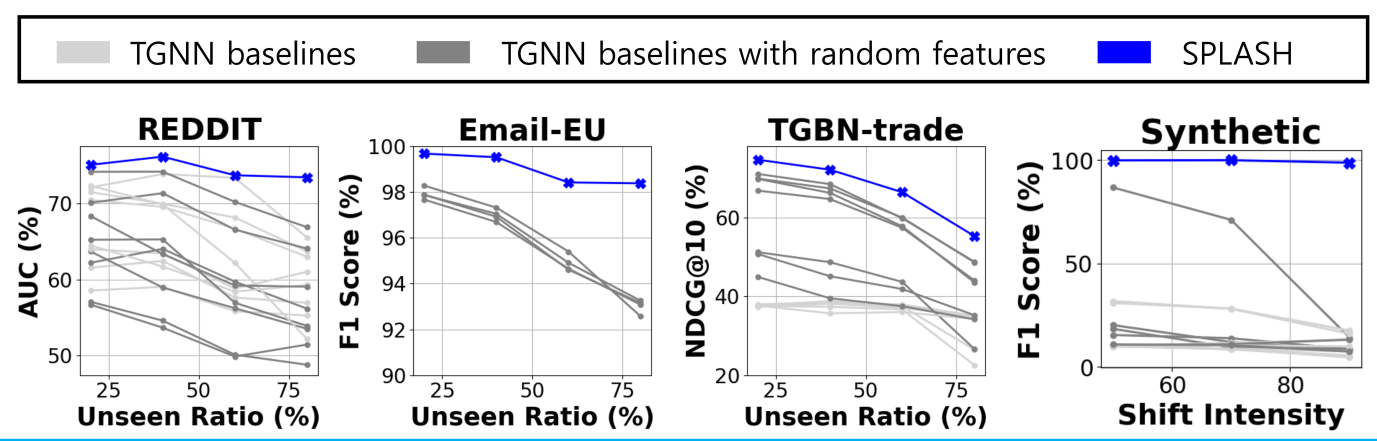


Experimental Results

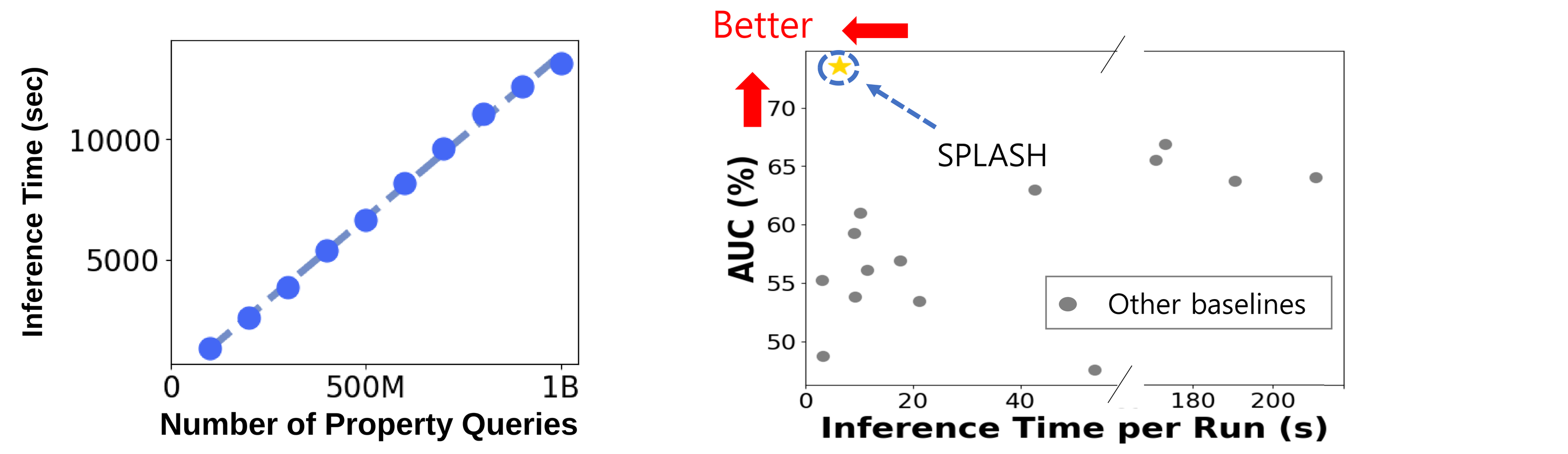
- **Research Questions**
  - We review our experiments for answering the following research questions:
    1) Accuracy & Generalization, 2) Efficiency & Scalability, 3) Qualitative Analysis.
- **RQ1) Accuracy & Generalization**
  - ***SPLASH outperforms all baselines in node property prediction, including TGNNs without node features and TGNNs using random features.***



  - ***SPLASH shows the least performance drop as the distribution shift increases, demonstrating better generalization capabilities compared to baselines.***



- **RQ2) Efficiency & Scalability**
  - ***SPLASH maintains a constant inference time per property query and offers the best trade-off between performance and inference speed.***



- **RQ3) Qualitative Analysis**
  - ***In the Email-EU dataset, which has static class properties, SPLASH generates the most distinct clusters for each class, compared to baselines.***